DOCUMENT RESUME

ED 336 949 FL 019 392

AUTHOR Griffin, Patrick E.; And Others
TITLE The Use of Latent Trait Methods To Examine Second Language Proficiency.
PUB DATE 91
NOTE 19p.
PUB TYPE Journal Articles (080)
JOURNAL CIT New Horizons: Hong Kong Teachers' Association Journal of Education; n26 p46-62 Nov 1985

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS Adults; Educational Objectives; *English (Second Language); Foreign Countries; Interviews; *Language Proficiency; Language Skills; *Language Tests; *Migrants; Second Language Instruction; Second Language Learning; *Student Evaluation; Test Construction; Test Items; *Test Reliability
IDENTIFIERS Australia (Victoria)

ABSTRACT

The Adult Migrant Education Service (AMES) of Victoria, Australia provides courses in English as a Second Language to non-English speaking migrants. Reviews currently under way are attempting to determine the effectiveness of this program and to find ways that might be used to help teachers assess the development of the students and to diagnose the difficulties so that instruction might be more effective. This study focuses on the development of instruments to assess client language skills. The diversity of ethnic groups and the wide range of client needs, background, personal characteristics, and courses offered made this a difficult task. This report describes the current assessment procedures of AMES, definitions of student language proficiency and achievement, definitions of language teaching objectives, the expansion of these objectives into test items, development of an interview test, field testing the interview test model, and analysis of student performance on the test. It was concluded that the reliability of the test was adequate for the purposes of individual diagnosis, and that the practice of embedding test items in a conversational flow enabled discourse to be assessed in a setting that simulated authentic conversation. (JL)

THE USE OF LATENT TRAIT METHODS TO EXAMINE SECOND LANGUAGE PROFICIENCY

by
Patrick E. Griffin, Raymond J. Adams,
Lynette Martin, and Barry Tomlinson
Education Department of Victoria, Australia

Introduction

The Adult Migrant Education Service (AMES) of Victoria, Australia provides courses in English as a Second Language (ESL) to non-English speaking migrants. Reviews currently under way are attempting to determine the effectiveness of this program and to find ways that might be used to assist teachers to assess the development of the students and to diagnose difficulties so that instruction might be more effective. This study focuses on the development of instruments to assess client language skills. However the diversity of ethnic groups, the wide range of client needs, background, personal characteristics and courses offered, made this a complex task.

The AMES provides a range of courses in different centres and the content of each course is determined by the needs of the student group undertaking the specific course. Consequently it is not possible to use content specific objectives as the basis of development if the instrument is to be used for different courses and centres. The language tasks used for assessment purposes have to be generic and enable data to be collected across centres and courses.

Current Assessment Procedures

At present, extensive use is made of the oral interview in placement procedures. In seeking admission to classes at the AMES, learners are interviewed by experienced teachers who globally assess the student's proficiency in the four macro-skills: listening, speaking, reading and writing. The student is then assigned to a class level based on the classification levels of the Australian Second Language Proficiency Ratings (ASLPR) developed from the Foreign Service Institute (FSI) rating scale by Ingram (1984). A short description of the

0 **Zero Proficiency**
Unable to function in the spoken language

0+ **Initial Proficiency**
Able to operate only in very predictable areas of need –
Vocabulary limited to that necessary to express simple elementary needs and basic courtesy formulae. Utterances generally consist of isolated words or short formulae.

1– **Elementary Proficiency**
Able to satisfy immediate needs using learned utterances –
First sign of spontaneity and flexibility emerging but there is no real autonomy of expression. Utterances generally made with fragmentary grammar, which may consist of no more than noun, verb, modifier.

1 **Minimum Survival Proficiency**
Able to satisfy basic survival needs and minimum courtesy requirements –
Can initiate and respond to simple statements and maintain very simple conversation within areas of immediate need or on very familiar topics. Fractured sentence structure and frequent grammatical errors. Can express likes and dislikes in areas of particular interest; can make basic survival transactions.

1+ **Survival Proficiency**
Able to satisfy survival needs and limited social demands –
Shows some spontaneity in language production but fluency uneven. Common tense forms occur and basic word order established. Can cope with less routine transactions in public. Some creativity with the language emerges.

2 **Minimum Social Proficiency**
Able to satisfy routine social demands and limited work requirements –
Can handle with confidence most social situations. Hesitations are still frequent but has vocabulary and grammar sufficient to express simply on most topics pertinent to everyday life. Does not have thorough control of longer grammatical constructions.

Figure 1 Summary of the ASLPR

first six levels of the ASLPR speaking are shown in Figure 1. Similar descriptions exist for each of the four macro skills. The ratings of the student's speaking and listening skills are generally based on informal discussion centered around personal information about the student's background and do not entail any strict form of testing. It was decided that this area should be the focus of the project, and to restrict the instrument to the assessment of proficiency in oral/aural language using an oral interview and a rating scale scoring procedure as proposed by Griffin (1985).

The ASLPR has only six categories (relevant to AMES) in which students can be rated. This is particularly blunt especially when it is considered that Ingram (1984) estimates that average students take about 240 hours to move from one level to another yet the longest AMES courses run for about 150-200 hours. The refinement of the accuracy of this process in the determination of the oral proficiency of the clients became the starting point of this project.

To identify sources for objective material and testing techniques a thorough examination of current AMES resources was undertaken. Of particular interest was the observation of classroom lessons. These observations highlighted the use of a wide range of methodologies and content areas.

Observations of classroom lessons also revealed that verbal assessment techniques used in the classroom consisted mainly of recitation exercises and a verbal equivalent of the cloze or sentence completion procedures. The students were prompted or even given the answer in the eliciting language. The difficulties associated with classroom oral/aural assessment did not appear to have been overcome and this impression was reinforced during a series of assessment workshops organised for teachers as part of the project.

Intuitive judgements were made, largely based on the experience of the teacher. For the less experienced teacher there did not seem to be any systematic way of monitoring achievement or developing proficiency. Because of the diverse nature of the courses, the only substantial data available on development was an end-of-course ASLPR classification. This made further exploration of proficiency rather than achievement an indispensible focus of this project.

The differences observed in classroom methodologies, the variety of courses, the variety of techniques and contexts used meant that a set of achievement or specific skill-based objectives would not be appropriate. Consequently it was decided that the objectives should be based on the notion of proficiency. Although the classes and resource material observed looked fundamentally different on the surface they all aimed to provide students with 'more English' so that they could more successfully communicate and function in 'real-life' situations.

Proficiency and Achievement

Numerous definitions and discussions of proficiency and communicative competence from a number of perspectives have appeared in the literature over the past decade (for example Canale and Swain, 1980; Hughes and Porter, 1983; Oller, 1983; Higgs, 1984; James, 1985; Rivera, 1985). For the purpose of this study the key feature of proficiency is its independence from curriculum and methodology. The distinction between achievement and proficiency is particularly significant. Fundamentally tests designed to measure these two phenomena differ according to the kind of information they are intended to supply.

Achievement tests measure the level of acquisition of specific course content. They are limited in scope and provide information about the extent to which a student has mastered particular material. On the other hand, proficiency testing assesses a student's language performance in terms of the extent to which language is used effectively outside the body of material specifically taught in class. The proficiency measure should represent the status of the student in terms of possible language use at that particular time. Proficiency testing then should be curriculum free and is not concerned directly with where, how, when or over what period of time the student developed the level of competence shown. The test should sample language tasks independent of the specific instructional material. In this manner the proficiency test should find the limits of language beyond which the student is unable to function at the time of testing.

To develop a set of proficiency based language tasks, an 'organizing principle' (Higgs, 1984) needs to be established. This proved to be particularly difficult and controversial because methodologies

and theories of language testing and teaching come and go rapidly.

In about 1972 Wilkins introduced what has become known as the functional/notional approach. The notional category consisted of ideas such as "frequency, duration, quantity, etc.". These items of meaning relate fairly directly to grammatical categories in European languages. Communicative functions are the broad uses to which we put language to express such uses as "requesting information", "giving orders" and so on. For many language teachers the functional/notional methodology offered a new approach and a large proportion of the language teaching community adopted it, thereby using it as a replacement for an organizing principle based on a traditional hierarchy of structural elements.

The definition of language teaching objectives in these terms, however, was not without difficulties. Lists of functions and notions were seen to be arbitrary and to preclude the generation of new sentences by the user. Whilst the ordering of skills and tasks is usually determined by a cross-fertilization between functional and grammatical categories, the generative grammatical system is thought to be fundamental (Brumfit, 1981). Hence the functional/notional approach alone did not appear useful as an 'organizing principle' for the development of an ordered sequence of functional language tasks. It did not allow for any assurance that the ordering would serve for most of the diverse courses offered.

Higgs and Clifford (1982) argued that it is inappropriate to specify any one element of language development as the overriding influence. In demonstrating the interrelationship among the elements of language development they developed a relative contribution model. This model incorporates most of the accepted notions of proficiency and indicates that any generic model should follow a particular sequence of development. The relative contribution of specific language sub-skills – pronunciation, vocabulary, grammar, fluency and sociolinguistic appropriateness changes from level to level. Relationships between these elements as proposed by Higgs and Clifford are illustrated in Figure 2.

Definitions and descriptions of developments in pronunciation, vocabulary, grammar, fluency and sociolinguistic skills are given in Liskin-Gasparro (1982). Basically pronunciation ranges from unintelligible to the fully acceptable pronunciation of an educated

native speaker. Fluency refers to the ease of production incorporating a range progressing from strained communication (except for routine expressions) to an ability to paraphrase with few fillers. Sociolinguistic appropriateness involves knowledge of how to deal with socio-cultural issues without offence, and the appropriate use and understanding of cultural references and expressions. Grammar is concerned with mastery of the structures of the language and ranges from the use of elementary constructions, through to the use of complete structures with low frequency of errors. Vocabulary develops from no knowledge of the new lexis through the use of a few isolated words and formulae that can be used to satisfy minimum everyday requirements to the use of sufficient vocabulary to engage in conversation and express opinions in formal and informal situations about a range of topics.

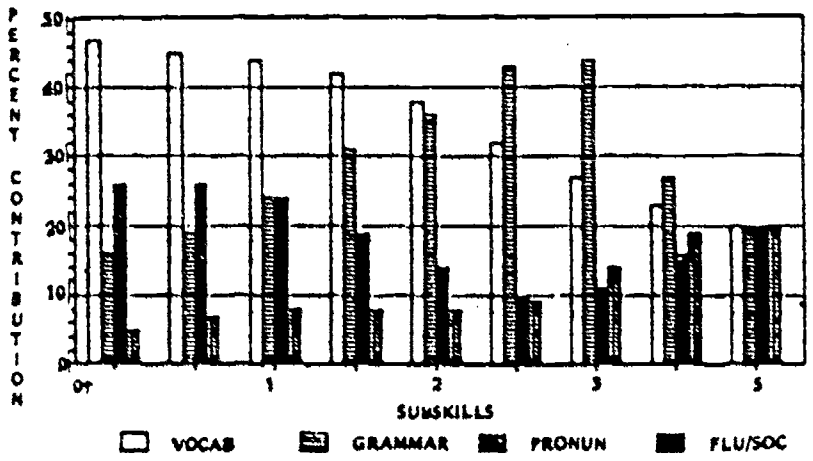

Figure 2 Contributions to Language Proficiency

According to this model vocabulary acquisition and pronunciation skills are the most important at the lowest levels of proficiency. This is consistent with the ASLPR and ACTFL/ETS descriptions (Liskin-Gasparro, 1982) which describe the acquisition of a limited vocabulary and a set of formulaic expressions as the first step. Beyond about level 1 the relative importance of vocabulary and pronunciation begins to drop and the importance of grammar rises rapidly. The factor that distinguishes between learners at about 1 to 2+ is the level of their

grammatical skill. It is here that the students begin to create with the language by using their grammatical skills to put the words and formulae together in useful structures. Beyond about 2+ the learners' sociolinguistic and fluency skills become the major area of improvement. As the ease of production increases, students begin to have register sensitivity and develop sociocultural awareness.

Since an upper limit of level 2 is used at the AMES this model would indicate that the acquisition of the grammatical system along with basic vocabulary would best serve as the organizing principle for proficiency objectives. However, traditional grammatical models, taught as a series of self-justifying rules and procedures have not only been criticised as stultifying but recent work (e.g., Johnston, 1985) even questions the validity of the ordering of these particular grammatical models.

Johnston (1985) analysed transcripts of unstructured naturalistic conversations for the emergence of certain morpho-syntactic structures and information on lexis. The analysis was restricted to the area of morpho-syntax, and only a small sample of Vietnamese and Polish immigrants was used. Nevertheless the results may offer a grammatical model of emerging linguistic forms. This information is worth considering as a complement to models proposed by others working from different premises. Although different learners in Johnston's sample showed individual variations, a continuum of increasing proficiency was evident. Johnston described the existence of "implicational relationships" that enable a prediction to be made about the structures a particular learner can or cannot be expected to use given his/her level of proficiency.

In developing oral proficiency tasks it seemed appropriate therefore to work from a structural approach which was functionally based, and to incorporate input from the work of Johnston, Higgs and Ingram.

There was also the problem of the communicative dimension. This must be considered an essential component in the development of objectives but the problem is how to combine the grammar, the functions and notions and this dimension into a workable model. According to the work of Higgs and Clifford the role of this dimension in the area of language testing in which this project was conducted is not so important as to play a major role in test development.

Defining Objectives

The ideas expressed above were then incorporated into a set of 33 amplified objectives in the style developed by Popham (1978). A sample amplified objective is provided in Figure 3. The objective is written in a form encompassing the ACTFL functional trisection (Liskin-Gasparro, 1982) which specifies its function, structure and context. Each amplified objective should therefore expand into a series of related items to be used as a criterion referenced test. The expansion of these objectives into test items should supply data with which to examine the implicational relationships identified by Johnston and might suggest reasons for the particular order of acquisition.

At this stage the item specifications were used to develop test items for an interview test. The set of objectives was grouped by structural type and the objectives were rank ordered by AMES teachers according the the most likely order of introduction into the classroom. As a group, the objectives cover up to five intended subtests ranging (in ASLPR terms) from 0+ to 2 and, as such, emphasise the use of vocabulary and grammatical structures.

Field Testing the Model

A pilot interview test was developed to examine the feasibility of applying the partial credit model (Wright and Masters, 1982). The interview was designed so that it could flow like a normal conversation. However, the conversation was carefully directed by the interviewer with each key item asked in a precise manner. Interspersed among the test items were related conversational points and hence not all responses were used for scoring purposes.

The interview and measurement model were trialed using 60 students previously rated as having ASLPR levels of 0+ to 1+ (i.e., 'initial proficiency' to 'survival proficiency'). Each interview was recorded and later scored by two raters. The scoring system that was used for each item is shown in Figure 4. It is presented as one possible example to show how samples of authentic language may be rated according to their degree of acceptability. The criteria shown in Figure 4 were applied to all of the test items. The aims in the development of these particular criteria were to provide easy to use rules that could be applied objectively in assessing structural accuracy, fluency and

appropriacy and while reflecting the lexical and grammatical basis of the items. In this study the same criteria were used for all items but the results suggest that in future it may be necessary to develop specific scoring criteria for each item in the test. This is particularly true if the implications of Johnston's and Higgs' results are to be incorporated.

Function:	Describe habitual actions, give information about other people.
Grammar:	Simple present tense (Johnson Verb 3SG-S level 5).
Response type:	Simple narrative.
Context:	Talking about regular activities of people.
Technique:	Open ended questions.
General Form:	What |does| |any subject| do |every| |time period| |do | |on | |specific time|
Example:	What does he do every day? What do you do on Saturdays?
Prop:	Series of pictures of a man going through a daily routine 1) Getting out of bed 2) Eating breakfast 3) Going into a building (factory) 4) Writing at desk 5) Buying a paper 6) Getting on the bus 7) Playing soccer 8) Watching T.V. 9) Washing face
Instruction:	Every day this man does these things, Point to the first picture and say:– First, he gets out of bed. Now tell me what he does. Point to each picture in turn. Then ask what do you do every day?

Figure 3 A Sample Objective

Score	Criteria
3	Appropriate information conveyed without errors, hesitation, self correction or prompts.
2	Appropriate information conveyed. One structural error may be present and/or prompts, hesitation and self correction.
1	Appropriate information conveyed but the response includes a number of structural errors. A considerable amount of prompting, hesitation and self correction may be included in the response.
0	Response not comprehensible, or not appropriate to the question asked.

Figure 4 Pilot Test Scoring Criteria

Results

Preliminary reliability studies indicated that good inter-rater agreement could be achieved if this scoring scheme was strictly applied. Each interview was scored by two raters. The regression of pairs of rating scores showed consistency of rating provided. This was evidenced by a regression slope of 1.00 with few points outside the 95% confidence bands. Internal consistency estimates of the reliability for the subtests varied between 0.85 and 0.95.

The data were analyzed using the CREDIT computer program (Masters, Wright and Ludlow, 1981). Since the test was divided into five separate units and many of the weaker students were not administered all of the units. Each unit was analyzed separately with CREDIT and then a common person equating procedure was used to transform all items on to a common scale (see Masters, 1984). Prior to equating, an examination of the item and person fit to the model was undertaken. Each item and person have an associated index of their fit to the model. If an item is found to have unacceptable fit, the hypothesis that it tests ability on the same proficiency dimension as the other items is rejected and the item was excluded from further analyses. For the pilot data the examination of fit led to the rejection of 10 items from a total of 64.

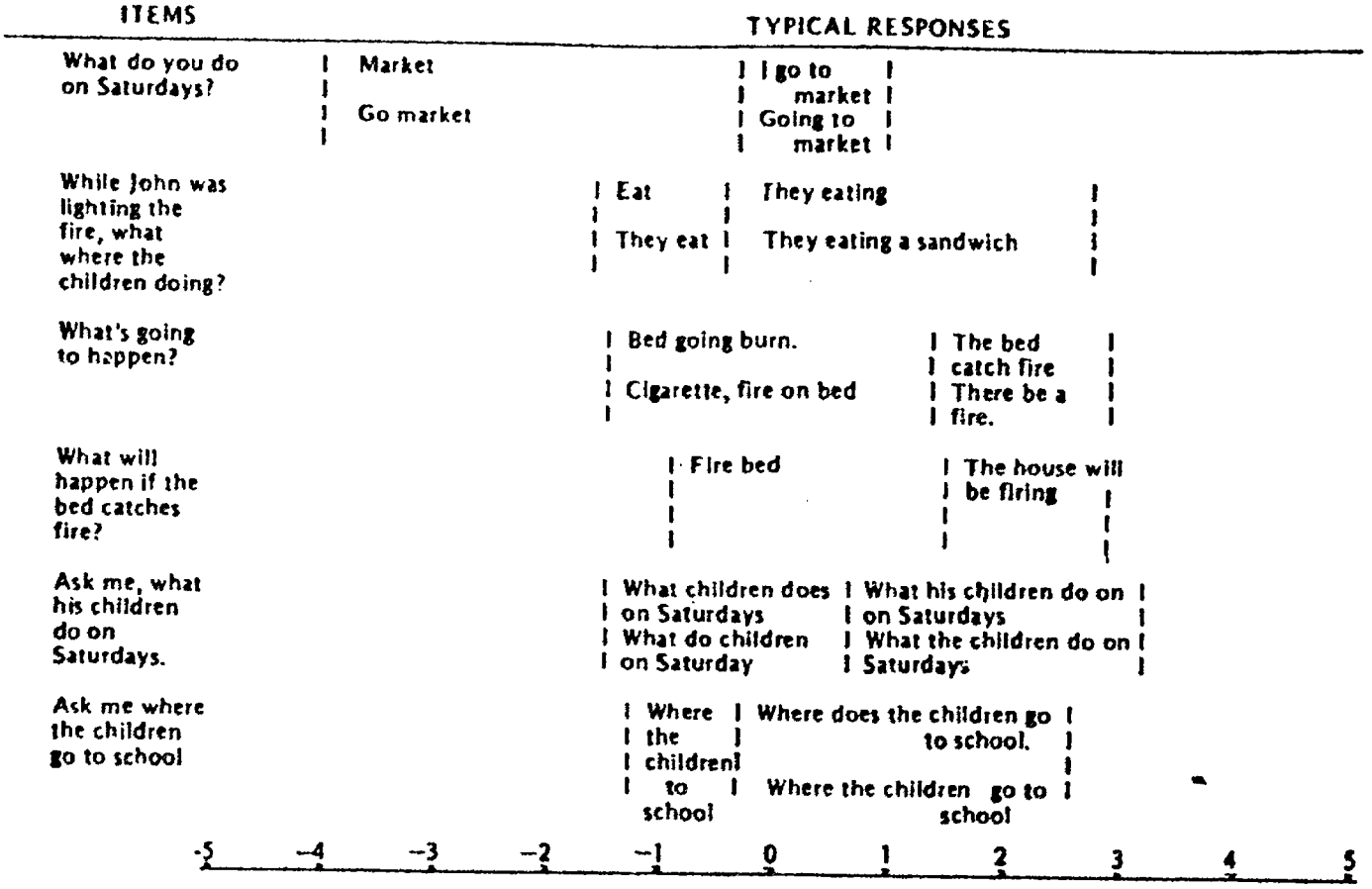

Figure 5 Map of the Proficiency Dimension

Since the data for the remaining 54 items could be shown to fit the requirements of the model the hypothesis that a second language proficiency dimension exists among the AMES students could not be rejected. Further, since the items were selected in order measure a portion of an assumed proficiency dimension among language tasks, the dimension was assumed to exist and its nature defined by mapping the items along that dimension in terms of their difficulty. An examination of typical incorrect responses made by students at various levels of proficiency might also provide the teacher with information needed for instructional purposes. Thus an interview planned as a conversation to provide some authenticity yields data which allows both student and teacher needs to be identified.

In Figure 5 a small sample of items has been mapped out. The scale values shown on the horizontal axis in Figure 5 describe the units of the proficiency dimension produced by the model. This scale is not the same as that used in the ASLPR, FSI or ACTFL/ETS materials. It is however an interval scale with proficiency increasing to the right and decreasing to the left. The advantage of the scale is that it is empirically based and, like the scales used in the ASLPR and other classification schemes, it is common to both the items and persons. Like those scales, it enables direct comparisons between student ability and language task difficulty. However the fineness of measurement allows greater precision than the classification schemes.

The score associated with a particular response is shown by its location on the map. For example, for the question 'What's going to happen?', the response, 'Bed going burn' scored one point, and a response worth one point was most likely for students with a proficiency between about −1.5 and 1.5 proficiency units. Students between about 1.5 and 2.8 proficiency units were most likely to score two with a response like 'There be a fire'. Students above 2.8 proficiency units were most likely to make a fully appropriate response and score three, while students below −2 were most likely to score zero.

Rejected Formats

Further analyses of the characteristic curves for individual items provide sufficient information to enable a thorough examination of the item's difficulty and the effectiveness of its scoring categories. This

includes a detailed analysis of the item discrimination over the proficiency range. For example, an analysis of the characteristic curves of two items that appeared to be almost identical indicates an apparent substantial difference in the difficulty.

Figure 6 shows the item characteristic curves for four of the items in the original test. The two items in Figures 6(i) and 6(ii) are not presented in Figure 5 and were eventually rejected from the test. The item in Figure 6(i), "Ask me what this man's name is" and the item in Figure 6(ii), "Ask me what this woman's name is" are almost identical but an inspection of the Figure shows that the item in 6(ii) was easier than the item in 6(i). For example a person of proficiency level about −2.2 is most likely to score a zero on the item in 6(i) and a one on the item in 6(ii). This may have been due to the unfamiliarity of the "Ask me" format in the item in 6(i) (the first occurrence of this). The same phenomenon is evident in other rejected items using the same format.

Figure 6(iii) illustrates the response curves of an example of an item where the "one" category is not dominant over any more than a very narrow proficiency range (the score of one indicating a low level but communicable response). When asked by the interviewer "Ask me where the children go to school" students over a wide proficiency range responded with "Where the children go to school?". In this case they are only echoing a section of the interviewer's question. Because of the frequent occurrence of the echo-type response it was felt that the item was unsuitable for inclusion in the test.

The item response curves in Figure 6(iv) illustrate an example of an item where both of the middle score categories were ineffective. Students either managed to make a fully appropriate response or showed no recognition of what was required. This item, "Can you tell me what they saw", was generally found to elicit a list. If the students recognised the requirements of the task they scored full marks and if not, zero. This item is essentially dichotomous and the use of the middle categories is artificial. This suggests that items which elicit lists or single words should be scored dichotomously rather than with the partial credit approach.

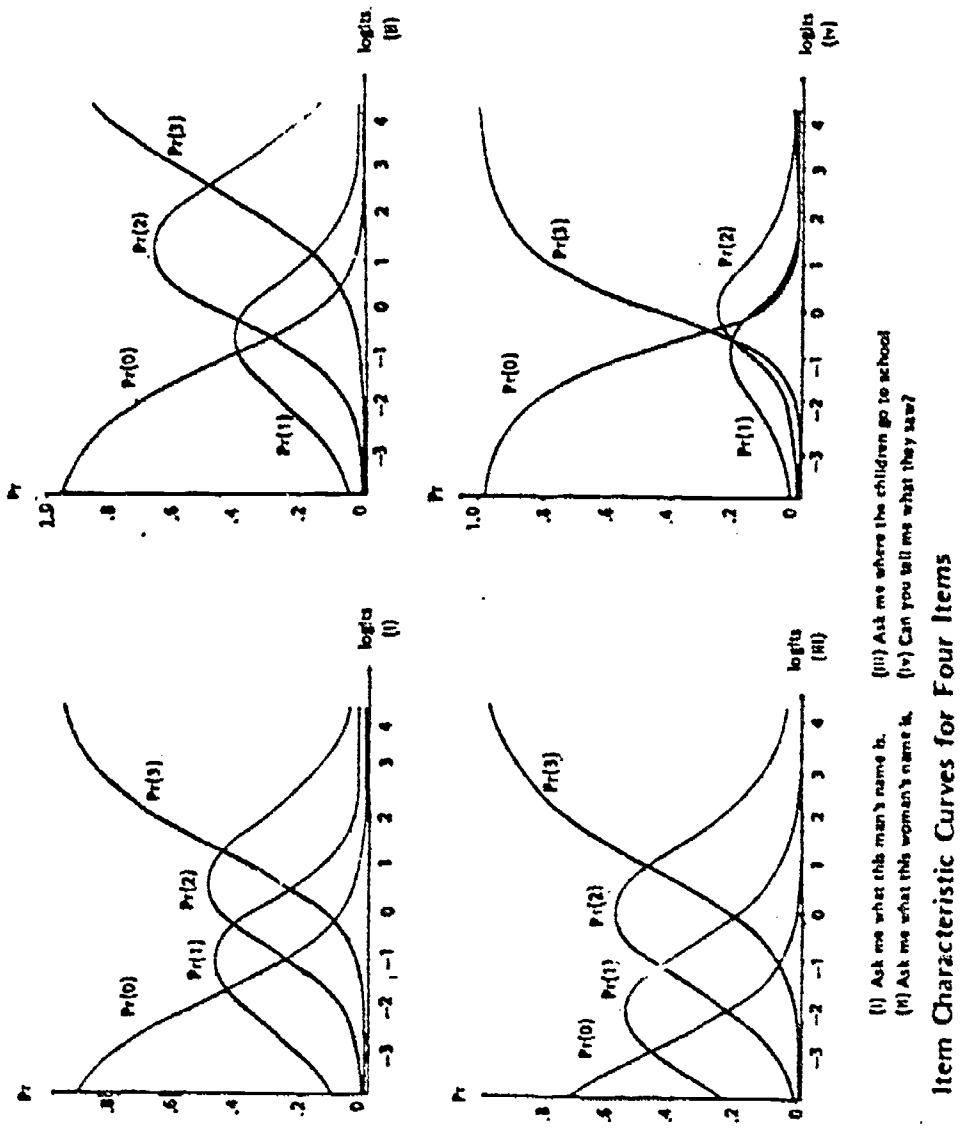

Figure 6 Item Characteristic Curves for Four Items

Analysis of Student Performance

Two statistics describing student performances are automatically produced by the application of Rasch model analyses: (1) an ability estimate, or the location of the student on the proficiency dimension and (2) the fit of the student response pattern to the model. A lack of fit indicates that an interpretation of the proficiency location for that student may require additional information related to student background, the curriculum or perhaps even the teaching method. The calibration is based on the expected response pattern of a group and, where an individual student response pattern does not deviate excessively from the group pattern, the individual is said to fit the model. Lack of fit generally occurs when a student fails to score as expected on an item or a set of items that are well above or below his/her proficiency level. For example, one student with a high overall proficiency had difficulties with the distinction between 'do' and 'does' and scored below expectation on items that involved these elements. Thus the lack of acceptable fit to the model (due to unexpected responses) can highlight specific persons with specific difficulties. In large scale testing and placement the use of the fit statistic has clear implications for the provision of diagonstic information regarding the relative strengths and weaknesses of an individual student.

Summary and Conclusions

By drawing on many of the accepted notions of proficiency it appears possible to develop a set of generic objectives that can be applied to a range of teaching methodologies and contexts.

When tests were developed from the objectives the general notion of an underlying dimension of proficiency could not be rejected, even from the limited data in this study. The data used for this paper covers only a small range of this dimension and in further investigation it will be essential to extend the dimension by using more difficult and varied communication tasks.

The linkage of the test units along the same dimension is perhaps the most encouraging aspect of the study. It indicates that a series of tests is likely to be measuring the same trait at different levels of proficiency. As a consequence, it seems possible to build a set of interview instruments which, when systematically scored, can provide a continuum of language proficiency development.

The reliability of the test was adequate for the purposes of individual diagnosis. The practice of embedding test items in a conversational flow enabled the discourse to be assessed in a setting which simulated authentic conversation. This, together with the overall fit of the data to a single dimensional model enhances the validity.

There were problems with the pilot instruments. The use of the question inversion type stimuli (i.e., "Ask me") tended to reduce the spread of responses. The question type and/or the criteria used had an artificial and levelling component. These types of questions coupled with the criteria tested in this trial should be avoided in the future. The use of a consistent rating scheme is unlikely to be appropriate for all items and future work will focus on the identification and development of separate scoring criteria for each objective. For example, in the case of simple language tasks requiring a list or a single word response, a dichotomous score may be required rather than the four categories used above. This introduces difficulties in scaling due to self weighting problems, and such issues still have to be addressed.

The pilot study was designed to test procedures rather than to assess the test items themselves. Difficulties associated with practice effects, lack of local independence, and with item structure all contribute to the decreased efficiency of the instrument. However, it is now possible to proceed with some confidence to define the dimension according to increasing language development and, from this, to develop a series of interview instruments which should be capable of diagnosing individual difficulties for placing students in appropriate instructional settings, and assisting teachers to plan both curriculum and instruction.

Further, long interviews characteristic of placement procedures in language courses may not be necessary. Initial probes by the interviewer should be able to broadly assess the interviewee's level on the proficiency dimension. At this point, the interviewer need only use part of the overall series of tests, reducing the time of testing but increasing the accuracy and effectiveness of the exercise. Subsequent stages of this study are needed to address the development of the instrument its calibration and validation in terms of its ability to provide similar broad based information similar to that currently available from other ESL instruments. In these stages the incorporation of the Higgs and

Clifford (1982) and Johnston (1985) models will be important in the definition of tasks and in the development of scoring criteria and procedures.

References

Brumfit, C. (1981). *Communication in the Classroom*. Harlow Essex: Longman.

Canale, M and Swain, M. (1980). Theoretical Bases of Communicative Approaches to Second Language Teaching and Testing. *Applied Linguistics. 11(1)*, 1-47.

Griffin, P.E. (1985). The use of latent trait methods in the calibration of spoken language in large-scale selection-placement programs. In Y. Lee, A. Fok, R. Lord and G. Low (eds), *New Directions in Language Testing*. London: Pergamon.

Higgs, T.V. (ed) (1984). *Teaching for Proficiency, the Organizing Principle*, Lincolnwood, Illinois: National Textbook Company.

Higgs, T.V. and Clifford, R. (1982). The Push Towards Communication. In T.V. Higgs (ed), *Curriculum Competence and the Foreign Language Teacher*. Lincolnwood, Illinois: National Textbook Company.

Hughes, A. and Porter, D. (eds) (1983). *Current Developments in Language Testing*. London: Academic Press.

Ingram, D.E. (1984). *Australian Second Language Proficiency Ratings*. Canberra.: Australian Government Printing Service.

James, C.J. (ed) (1985). *Foreign Language Proficiency in the Classroom and Beyond*. Lincolnwood, Illinois: National Textbaook Company.

Johnston, M. (1985). *Syntactic and Morphological Progressions in Learner English*. Department of Immigration and Ethic Affairs.

Liskin-Gasparro, J.E. (1982). *ETS Oral Proficiency Testing Manual*. Princeton, N.J.: Educational Testing Service.

Masters, G.N. (1984). Constructing an item bank using partial credit scoring. *Journal of Educational Measurement, 21*, 19-32.

Masters, G.N., Wright, B.D., and Ludlow, L. (1981). CREDIT. A Rasch program for ordered response categories. MESA Psychological Laboratory. University of Chicago.

Oller, J.W. (ed) (1983). *Issues in Language Testing Research*. Rowley, Mass.: Newbury House.

Popham, W.J. (1978). *Criterion Referenced Measuremet*. Englewood Cliffs: Prentice Hall.

Wright, B.D. and Masters, G.N. (1982). *Rating Scale Analysis: Rasch Measurement*. Chicago: MESA Press.

NEW HORIZONS

CONTENTS

THE JOURNAL OF EDUCATION

HONG KONG TEACHERS' ASSOCIATION

NO. 26 NOVEMBER 1985